Ilmenauer Beiträge zur Wirtschaftsinformatik

Herausgegeben von U. Bankhofer, V. Nissen
D. Stelzer und S. Straßburger

Jürgen Vogel

Die Güte von Anpassungstests auf Benfordverteilung

Arbeitsbericht Nr. 2012-01, Juli 2012

Technische Universität Ilmenau
Fakultät für Wirtschaftswissenschaften
Institut für Wirtschaftsinformatik

Autor: Jürgen Vogel

Titel: Die Güte von Anpassungstests auf Benfordverteilung

Ilmenauer Beiträge zur Wirtschaftsinformatik Nr. 2012-01, Technische Universität Ilmenau, 2012

ISSN 1861-9223

ISBN 978-3-938940-42-6

urn:nbn:de:gbv:ilm1-2012200141

© 2012 Institut für Wirtschaftsinformatik, TU Ilmenau

Anschrift: Technische Universität Ilmenau, Fakultät für Wirtschaftswissenschaften, Institut für Wirtschaftsinformatik, PF 100565, D-98684 Ilmenau.
http://www.tu-ilmenau.de/wid/forschung/ilmenauer-beitraege-zur-wirtschaftsinformatik/

Gliederung

Zusammenfassung: *In umfangreichen empirischen Datensätzen erweist sich die erste Ziffer der dort aufgelisteten Dezimalzahlen häufig als nicht gleichmäßig verteilt, vielmehr genügt sie der sogenannten benfordschen Verteilung. Eine nachweisbare Abweichung von dieser Verteilung z. B. in Wirtschaftsbilanzen könnte ein Hinweis auf Manipulation sein. Daher ist es wichtig, ein gutes statistisches Verfahren zum signifikanten Nachweis der Abwesenheit von Benfordverteilung zu haben. Der dafür prädestinierte Chiquadrat-Anpassungstest wird hier mit sieben anderen aus der Literatur bekannten Tests, die zur Anpassung speziell an die Benfordverteilung entwickelt worden sind, bezüglich seiner Güte verglichen. Am besten schneidet in diesem Vergleich, der anhand von Monte-Carlo-Simulationen durchgeführt wird, ein modifizierter Watson-Test ab.*

Schlüsselwörter: *Benfordverteilung, Gesetz von Benford, Anpassungstests, Güte von Tests, Watson-Test*

1 Das Gesetz von Benford

Im Jahre 1938 hat der amerikanische Physiker Frank Benford [1] darauf hingewiesen, dass in vielen empirischen Datenreihen die aufgelisteten Dezimalzahlen häufiger mit einer „1“ beginnen als mit der Ziffer „2“, diese häufiger ist als die „3“ usw. Er hat für die Verteilung der ersten Ziffer eine Art logarithmische Verteilung angegeben (siehe Formel (B) auf der nächsten Seite), die seitdem mit seinem Namen verbunden ist, auch wenn wir heute wissen, dass der Astronom Simon Newcomb [10] dasselbe Gesetz schon 57 Jahre vorher veröffentlicht hatte. Beide Wissenschaftler sind zu ihrer Entdeckung durch die Beobachtung gelangt, dass die vorderen Seiten in Logarithmentafeln stärker abgenützt sind als die hinteren. Das benfordsche Gesetz kommt aber auch bei vielen anderen Datensätzen zum Tragen, wie z. B. bei der Größe von Gewässern, Ernteergebnissen, Hausnummern in Adressverzeichnissen, Preisen in einem Supermarkt, Einwohnerzahlen von Städten oder in umfangreichen Listen von Naturkonstanten.

Das benfordsche Gesetz ist inzwischen gründlich erforscht. Eine gute Beschreibung des mathematischen Hintergrundes kann man bei Berger und Hill [2] finden. Charakteristisch für die Benfordverteilung ist die Skaleninvarianz, das heißt, nach Multiplikation aller Daten mit einer Konstanten ändert sich die Verteilung der ersten Ziffer nicht. Diese Eigenschaft hat nur die Benfordverteilung. Damit erklärt sich auch das häufige Vorkommen des Phänomens in Finanzmarktdatensätzen, bei denen die Verteilung der ersten Ziffer sinnvollerweise nicht davon abhängen sollte, welche Währung benutzt wird. Eine andere Begründung für das Wirken des benfordschen Gesetzes ist ein Grenzwertsatz, nach dem das Produkt $X_1 \cdot X_2 \cdot ... \cdot X_n$ von unabhängigen, identisch nicht diskret verteilten Zufallsvariablen mit wachsendem n in Verteilung gegen eine Zufallsvariable konvergiert, deren erste Ziffer benfordverteilt ist ([2], Seite 104). Deswegen treten benfordverteilte erste Ziffern vorrangig in Zahlenfolgen auf, die einem natürlichen Wachstumsprozess unterliegen. Wenn ein konkreter Datensatz wider alle Erwartung nicht dem benfordschen Gesetz genügt, entsteht die Frage, ob die Daten vielleicht manipuliert worden sind. Tatsächlich hat man auf diese Weise schon Betrügereien bei Krankenkassen, in Steuererklärungen und in Unternehmensbilanzen aufgedeckt, wie z. B. in [3] und in [4] beschrieben wird. Zum statistischen Nachweis der Abwesenheit von Benfordverteilung benötigt man einen geeigneten Signifikanztest. Im Folgenden werden mehrere solcher Tests, die aus der Literatur be-

kannt und einfach durchzuführen sind, im Detail vorgestellt und bezüglich ihrer Güte miteinander verglichen.

2 Anpassungstests zum Prüfen auf Benfordverteilung

Signifikante Abweichungen empirischer Daten von einer theoretischen Verteilung werden in der Statistik mit einem Anpassungstest nachgewiesen. Dazu sei angenommen, dass zu einem metrisch skalierten Merkmal X eine reine Zufallsstichprobe vom Umfang n vorliegt und die Beobachtungswerte Dezimalzahlen sind. D_1 bezeichne die erste Ziffer des Merkmals X, also die führende Ziffer nach Weglassen des Vorzeichens, des Dezimaltrennzeichens und aller voranstehenden Nullen. Als theoretische Verteilung wird die diskrete Wahrscheinlichkeitsverteilung auf den ganzen Zahlen $k \in \{1,2,...,9\}$ mit den Einzelwahrscheinlichkeiten

$$P(D_1 = k) = b_k := \log_{10}\left(\frac{k+1}{k}\right) = \log_{10}(k+1) - \log_{10}(k) \qquad \text{(B)}$$

angenommen, die hier als *Benfordverteilung* bezeichnet wird[1]. In Abbildung 3 weiter hinten im 4. Abschnitt ist diese Verteilung zusammen mit zwei anderen Verteilungen als Wahrscheinlichkeitsdiagramm grafisch dargestellt.

Die zu überprüfende Nullhypothese lautet

„Die erste Ziffer in der Dezimaldarstellung des Merkmals X genügt der Verteilung (B)."

Als Anpassungstest für eine diskrete Verteilung gibt es eigentlich nur den pearsonschen Chiquadrat-Test, der in dem Ruf steht, nicht besonders mächtig zu sein. Andere bekannte Tests zum Prüfen auf Benfordverteilung sind Modifikationen von Anpassungstests, die ursprünglich für stetige Verteilungen entwickelt wurden. Verwendet man diese Tests für diskrete Verteilungen, begeht man in der Regel keinen Fehler, aber man verschenkt Signifikanz. Dieses als „konservativ" bezeichnete Verhalten bedeutet, dass de facto mit einem kleineren Signifikanzniveau gearbeitet wird, als man vorgegeben hat. Das wäre nicht weiter schlimm, wenn damit nicht auch falsche Nullhypothesen mit unnötig kleiner Wahrscheinlichkeit abgelehnt würden. Dadurch haben solche Tests a priori eine geringe Macht. Für Tests, die zur Anpassung an nur eine bestimmte diskrete Verteilung dienen

[1] In theoretischen Untersuchungen zum benfordschen Gesetz ist es auch üblich, als *Benfordverteilung* die stetige Wahrscheinlichkeitsverteilung mit der Verteilungsfunktion $F(x) = \log_{10} x \quad (1 \le x < 10)$ zu bezeichnen.

sollen, bietet sich als Ausweg an, den kritischen Bereich so zu vergrößern, dass keine Signifikanz mehr verschenkt wird. In diesem Sinne hat J. Morrow [9] für den Kolmogorow-Smirnow-Test (K-S-Test) und den mit ihm verwandten Kuiper-Test neue, asymptotisch gültige kritische Schranken speziell zur Prüfung auf Benfordverteilung bestimmt. Das Gleiche hatte vorher schon D. E. Giles [7] mit dem Watson-Test gemacht.

Diese Tests sollen hier auf ihre Güte untersucht werden. Außerdem werden in den Vergleich der Chiquadrat-Anpassungstest und zwei in [8] bzw. [4] vorgeschlagene Tests einbezogen. Die kritischen Schranken zu letzteren, dem Maximum- und dem Distanz-Test, sind ebenfalls von Morrow [9] berechnet worden. Allen hier behandelten Tests ist gemeinsam, dass die kritischen Schranken vom Stichprobenumfang unabhängig sind. Dadurch werden keine umfangreichen Tabellen mit Quantilen benötigt. Und weil sich auch die Testgrößen leicht mit einem Computer oder Taschenrechner berechnen lassen, ist die praktische Handhabung sehr einfach. Allerdings sind die kritischen Schranken in der Regel nur asymptotisch für $n \to \infty$ gültig, so dass es bei kleinen Stichprobenumfängen zu einer erheblichen Unterschreitungen des Signifikanzniveaus kommen kann. Um dies zu vermeiden, hat M. A. Stephens in [5] die Formeln zur Berechnung von Testgrößen, die auf der empirischen Verteilungsfunktion basieren, so verändert, dass die Nullhypothese bei kleinen Stichprobenumfängen eher verworfen wird. Zwei dieser modifizierten Tests, der K-S- und der Kuiper-Test, sind in den Gütevergleich mit aufgenommen worden.

Im Folgenden werden die Berechnungsvorschriften für die Testgrößen zusammengestellt. Es bezeichne f_k die relative Häufigkeit des Auftretens von k als erste Ziffer in der Stichprobe ($k = 1, 2, \ldots, 9$) und b_k die entsprechende Einzelwahrscheinlichkeit der Benfordverteilung gemäß (B). Die Stichprobenfunktion

$$T_j := \sum_{k=1}^{j} (f_k - b_k)$$

gibt die Abweichung zwischen der empirischen Verteilungsfunktion und der hypothetischen Verteilungsfunktion an der Stelle j an ($j = 1, 2, \ldots, 9$). Die beiden Tests vom K-S-Typ basieren auf den maximalen Differenzen

$$D^+ := \max_{j=1,\ldots,9}(T_j) \quad \text{und} \quad D^- := \max_{j=1,\ldots,9}(-T_j)$$

der beiden genannten Verteilungsfunktionen. Daraus werden dann die bekannten Stichprobenfunktionen

$$D := \max(D^+, D^-) \quad \text{und} \quad V := D^+ + D^-$$

für den K-S-Test bzw. den Kuiper-Test gebildet. Der Watson-Test verwendet die von L. S. Freedman [6] modifizierte Testgröße

$$U^2 = \frac{n}{9} \cdot \left[\sum_{j=1}^{8} T_j^2 - \frac{1}{9} \left(\sum_{j=1}^{8} T_j \right)^2 \right].$$

Die beiden anderen von Morrow vorgeschlagenen Tests verwenden die Maximaldistanz m und die euklidische Distanz d gemäß

$$m := \max_{j=1,\ldots,9} \left| f_j - b_j \right| \quad \text{bzw.} \quad d := \sqrt{\sum_{j=1}^{9} (f_j - b_j)^2},$$

während der Chiquadrat-Anpassungstest das pearsonsche Chiquadrat

$$\chi^2 := n \cdot \sum_{j=1}^{9} \frac{(f_j - b_j)^2}{b_j}$$

als Testgröße benutzt. Die für diesen Test empfohlene Voraussetzung $n \cdot b_k \geq 5$ für alle k ist bei der Benfordverteilung wegen $n \cdot b_9 \approx n \cdot 0{,}04825 \geq 5$ erst ab einem Stichprobenumfang von $n = 104$ erfüllt.

Testname	Teststatistik	Kritische Schranken $\alpha = 0{,}10$	$\alpha = 0{,}05$	$\alpha = 0{,}01$
K-S-Test	$\sqrt{n} \cdot D$	1,01	1,15	1,42
K-S-Test modifiziert	$\left(\sqrt{n} + 0{,}12 + \frac{0{,}11}{\sqrt{n}} \right) \cdot D$	1,012	1,148	1,420
Kuiper-Test	$\sqrt{n} \cdot V$	1,19	1,32	1,58
Kuiper-Test modifiziert	$\left(\sqrt{n} + 0{,}155 + \frac{0{,}24}{\sqrt{n}} \right) \cdot V$	1,191	1,321	1,579
Morrows Maximum-Test	$\sqrt{n} \cdot m$	0,851	0,967	1,212
Morrows Distanz-Test	$\sqrt{n} \cdot d$	1,212	1,330	1,569
Watson-Test modifiziert	U^2	0,14313	0,17878	0,26319
Chiquadrat-Test	χ^2	13,36	15,51	20,09

Tabelle 1: Testgrößen und kritische Schranken für die acht untersuchten Tests

In Tabelle 1 sind die Testgrößen der untersuchten Tests und die kritischen Schranken zu den Signifikanzniveaus 10 %, 5 % und 1 % zusammengestellt. Die kritischen Schranken für den Pearson-Test sind Quantile der Chiquadrat-Verteilung mit 8 Freiheitsgraden und die für den Watson-Test sind [7] entnommen worden. Die anderen Signifikanzgrenzen stammen von Morrow [9]. Die Nullhypothese ist immer dann zurückzuweisen, wenn der Wert der Testgröße die kritische Schranke übersteigt.

3 Design und Ergebnisse der Simulationsstudie

Bei der Güteuntersuchung von Signifikanztests geht es vor allem um die Wahrscheinlichkeit der Ablehnung der Nullhypothese. Wenn die Nullhypothese wahr ist, darf die Wahrscheinlichkeit ihrer Rückweisung nicht größer als das vorgegebene Signifikanzniveau α sein. Eine falsche Nullhypothese sollte dagegen mit möglichst großer Wahrscheinlichkeit zurückgewiesen werden. Diese als Macht bezeichnete Wahrscheinlichkeit hängt nicht nur vom Stichprobenumfang, sondern auch davon ab, was konkret mit „falscher Nullhypothese“ gemeint ist. Wenn die Beobachtungswerte aus einer Grundgesamtheit stammen, die nicht benfordverteilt ist, so kann sich je nach Verteilung einmal der eine, ein anderes Mal der andere Test als mächtiger erweisen. Nur selten gelingt es einen Test zu konstruieren, der bei allen Arten der Abweichung von der Nullhypothese der empfindlichste ist.

Zur Güteuntersuchung der Tests wurden die Rückweisewahrscheinlichkeiten durch Monte-Carlo-Simulation näherungsweise bestimmt. Dazu sind für verschiedene Verteilungen jeweils 1 Million Datensätze mit den Stichprobenumfängen 20, 30, 40, 50, 75, 100, 125, 150, 200, 300, 400, 500 und 750 in Form von Pseudozufallszahlen erzeugt worden. Die erste Ziffer dieser Zahlen ist mit den Tests jeweils zum Signifikanzniveau 10 %, 5 % und 1 % auf Vorliegen von Benfordverteilung geprüft und die Häufigkeit der Ablehnung der Nullhypothese gezählt worden. Die Generierung der Zufallszahlen erfolgte mit dem Mersenne-Twister-Algorithmus in der Statistiksoftware R. Für jeden Stichprobenumfang und jedes Signifikanzniveau kamen gesonderte Datensätze zum Einsatz.

Zur Analyse der Irrtumswahrscheinlichkeit bei Rückweisung der (wahren) Nullhypothese sind benfordverteilte Daten erzeugt worden. Das geht einfach, indem man gleichmäßig auf [0, 1) verteilte Zahlen generiert und diese in den Exponenten von 10 setzt. Die Ergebnisse dieser Analyse sind in der Abbildungen 1 für die Tests vom K-S-Typ und in Abbildung 2 für die vier anderen Tests grafisch dargestellt.

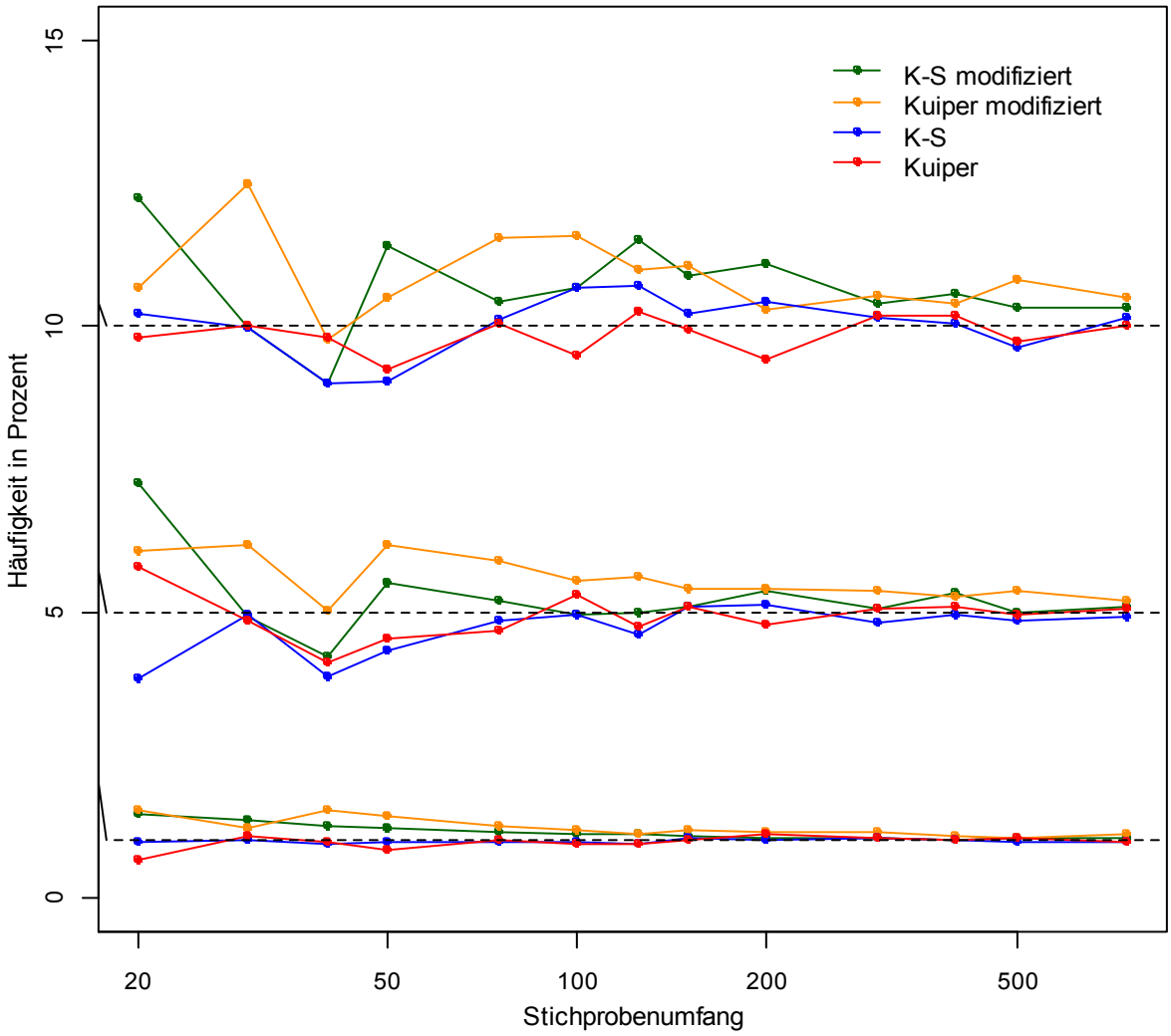

Abbildung 1: Relative Rückweisehäufigkeiten benfordverteilter Ziffern zu den Signifikanzniveaus 10 %, 5 % und 1 % durch Anpassungstests vom K-S-Typ

Zum Machtvergleich benötigt man nicht benfordverteilte Datensätze. Liegen Beobachtungswerte vor, deren erste Ziffer einer deutlich von Benford abweichenden Verteilung genügt, wie z. B. binomial verteilte, poissonverteilte oder gleichmäßig verteilte Daten, so weisen alle acht Tests selbst bei kleinen Stichprobenumfängen die Nullhypothese mit großer Wahrscheinlichkeit zurück. Umgekehrt gibt es z. B. bei mit dem Parameter $\sigma = 1$ lognormalverteilten Daten, deren erste Ziffer nahezu benfordsch ist, ebenfalls keine Unterschiede zwischen den Tests, weil bei allen die Rückweisehäufigkeit genau so klein ist wie bei der Benfordverteilung selbst. Es werden also nicht benfordverteilte Datensätze gebraucht, deren Verteilung nicht zu stark von der benfordschen abweicht.

Exemplarisch dafür werden hier die Normalverteilung mit $\mu = 0$ und $\sigma = 0{,}6$ und die Lognormalverteilung mit $\mu = 0$ und $\sigma = 0{,}7$ zum Gütevergleich herangezogen. Sie gehören zu

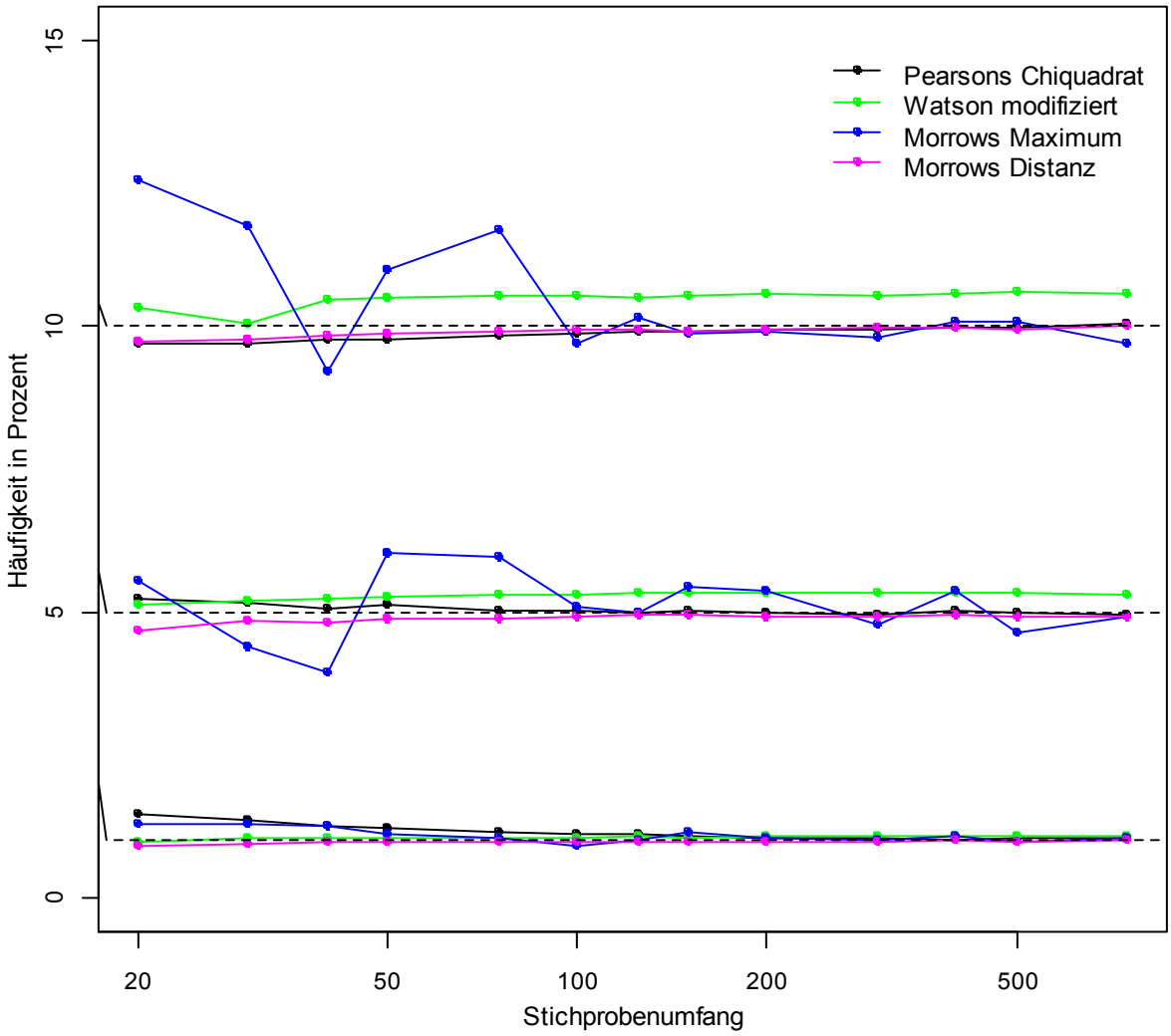

Abbildung 2: Relative Rückweisehäufigkeiten benfordverteilter Ziffern zu den Signifikanzniveaus 10 %, 5 % und 1 % durch vier weitere Anpassungstests

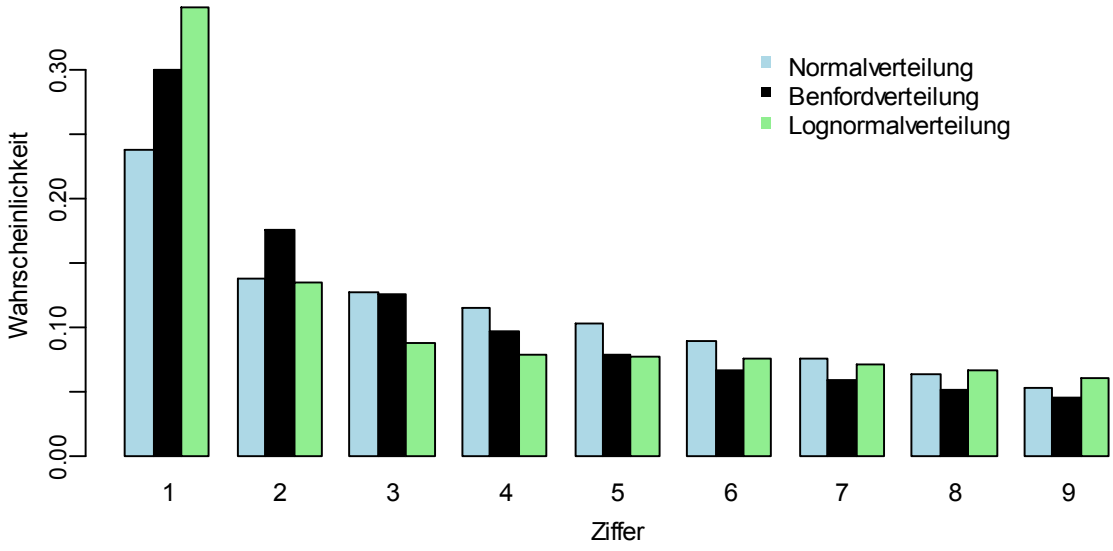

Abbildung 3: Verteilung der ersten Ziffer von N(0; 0,6²)- und LN(0; 0,7²)-verteilten Zufallsvariablen im Vergleich zur Benfordverteilung

zwei Verteilungstypen, die in der Realität häufig vorkommen und gerade bei Wirtschafts- und Finanzdaten eine praktisch relevante Alternative zur Benfordverteilung sein könnten. Der Unterschied dieser zwei Verteilungen beim Erste-Ziffer-Vergleich mit der Benfordverteilung wird in Abbildung 3 sichtbar gemacht. Die Ergebnisse der Analyse sind dann für

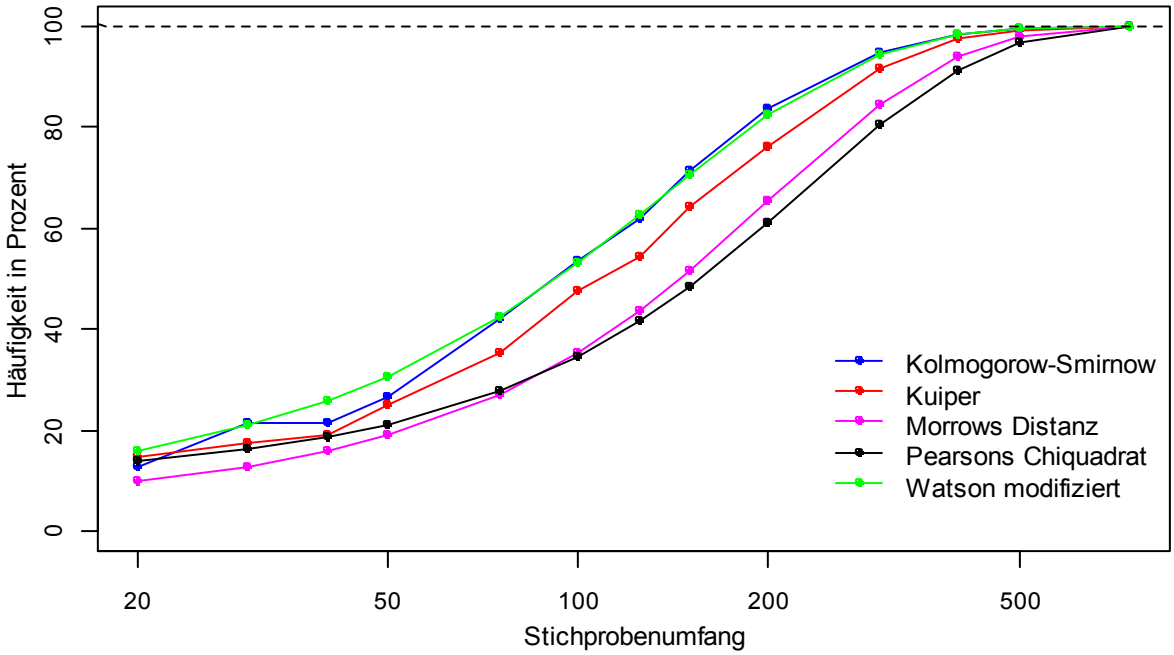

Abbildung 4: Relative Rückweisehäufigkeiten für $\alpha = 0{,}05$ bei N(0; 0,6²)-verteiltem Merkmal

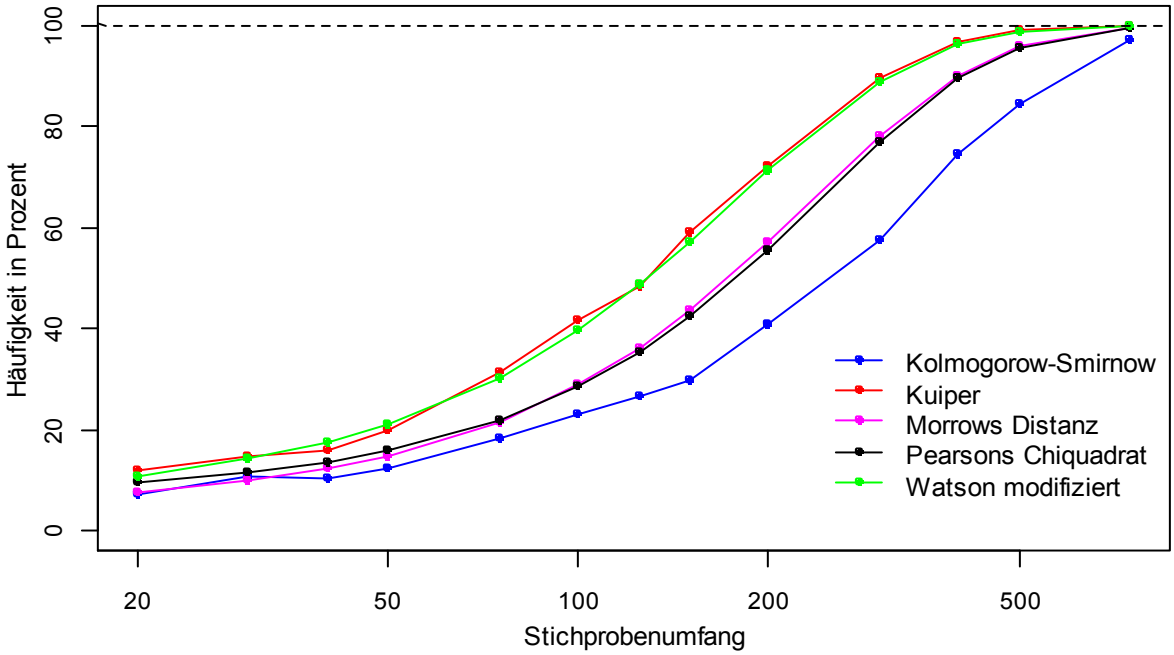

Abbildung 5: Relative Rückweisehäufigkeiten für $\alpha = 0{,}05$ bei LN(0; 0,7²)-verteiltem Merkmal

$\alpha = 0{,}05$ in den Abbildungen 4 und 5 grafisch dargestellt. Für die beiden anderen Niveaus $\alpha = 0{,}01$ und $\alpha = 0{,}10$ fallen die Ergebnisse, die hier nicht präsentiert werden, völlig analog aus.

4 Schlussfolgerungen und Fazit

Beim Betrachten der Rückweisehäufigkeiten in den Abbildungen 1 und 2 fällt auf, dass der modifizierte K-S-Test, der modifizierte Kuiper-Test und Morrows Maximum-Test das vorgegebene Signifikanzniveau α nicht einhalten. Nicht nur bei kleinen Stichprobenumfängen lehnen diese drei Tests die wahre Nullhypothese mit einer relativen Häufigkeit ab, die zum Teil deutlich über α liegt. Stephens' Modifizierung bringt hier eher Nachteile und sollte deshalb nicht verwendet werden. Sehr gut dagegen verhält sich diesbezüglich Morrows Distanz-Test, der nicht nur das Signifikanzniveau permanent einhält, sondern auch fast nichts verschenkt. Der Watson-Test liegt mit seiner Irrtumswahrscheinlichkeit beständig leicht über dem Signifikanzniveau, was man aber wegen seiner ansonsten hervorragenden Qualität hinnehmen sollte.

Die Güte der Tests lässt sich anhand der Abbildungen 4 und 5 vergleichen. Die drei Tests, die das Signifikanzniveau deutlich überschreiten, sind in dem Vergleich nicht mehr enthalten. Es sei hier nur angemerkt, dass Morrows Maximum-Test zu den weniger mächtigen Tests gehört. Sowohl die Normalverteilung als auch die Lognormalverteilung werden am besten vom Watson-Test als falsch erkannt. Er besitzt bei allen Stichprobenumfängen die größte Macht. Der K-S-Test ist bei der Normalverteilung und der Kuiper-Test bei der Lognormalverteilung dem Watson-Test gleichwertig. Während der Kuiper-Test auch im anderen Fall gut abschneidet, lehnt der K-S-Test die Lognormalverteilung mit der geringsten Wahrscheinlichkeit ab. Der Chiquadrat-Anpassungstest und Morrows Distanz-Test unterscheiden sich kaum in ihrer Macht, was bei der nahen Verwandtschaft ihrer Testgrößen nicht weiter verwundert. Sie sind insgesamt schlechter als der Watson- und der Kuiper-Test.

Als Fazit kann man folgende Empfehlung geben: Zur Prüfung auf Vorliegen von Benfordverteilung bei der ersten Ziffer empirischer Daten benutze man zunächst den von Giles [7] beschriebenen Watson-Test. Auch der (nicht modifizierte) Kuiper-Test mit den kritischen Schranken von Morrow [9] wäre eine gute Wahl. Führt keiner dieser Tests zur Ablehnung der Nullhypothese, kann die Anwendung des Kolmogorow-Smirnow- und des Chiquadrat-

Tests sinnvoll sein, weil sie bestimmte Verteilungen mit großer Wahrscheinlichkeit als nicht benfordsch zurückweisen. So hat sich Letzterer zwar bei den zwei hier verwendeten Alternativverteilungen als weniger mächtig erwiesen, er erkennt jedoch besser als die anderen Tests, wenn eine einzelne Ziffer im Vergleich zur Benfordverteilung besonders oft oder besonders selten an erster Stelle steht.

Literatur

[1] Benford, F. (1938): The law of anomalous numbers, Proceedings of the American Philosophical Society, 78, 551-572.

[2] Berger A.; Hill, T. P. (2011): A basic theory of Benford's Law. Probability Surveys, 8, 1-126.

[3] Rauch, B.; Göttsche, M.; Brähler, G.; Engel, S. (2011): Fact and fiction in EU-governmental economic data. German Economic Review, 12(3), 243-255.

[4] Cho, W. K. T.; Gaines, B. J. (2007): Breaking the (Benford) law: Statistical fraud detection in campaign finance. The American Statistician, 61, 218-223.

[5] D'Agostino, R. B.; Stephens, M. A. (1986): Goodness-of-Fit Techniques, Marcel Dekker.

[6] Freedman, L. S. (1981): Watson's U_N^2 statistic for a discrete distribution. Biometrika 68, 708-711.

[7] Giles, D. E. (2006): The exact asymptotic distribution function of Watson's U_N^2 for testing goodness-of-fit with circular discrete data. University of Victoria, Econometrics Working Paper EWP0607.

[8] Leemis, L. M.; Schmeiser, B. W.; Evans, D. L. (2000): Survival distributions satisfying Benford's law. The American Statistician, 54, 236-241.

[9] Morrow, J. (2010): Benford's law, families of distributions and a test basis. http://www.johnmorrow.info/projects/benford/benfordMain.pdf

[10] Newcomb, S. (1881): Note on the frequency of use of the different digits in natural numbers. American Journal of Mathematics, 4(1), 39-40.